

Separable covariance arrays via the Tucker product, with applications to multivariate relational data

Peter D. Hoff *

August 13, 2010

Abstract

Modern datasets are often in the form of matrices or arrays, potentially having correlations along each set of data indices. For example, data involving repeated measurements of several variables over time may exhibit temporal correlation as well as correlation among the variables. A possible model for matrix-valued data is the class of matrix normal distributions, which is parametrized by two covariance matrices, one for each index set of the data. In this article we describe an extension of the matrix normal model to accommodate multidimensional data arrays, or tensors. We generate a class of array normal distributions by applying a group of multilinear transformations to an array of independent standard normal random variables. The covariance structures of the resulting class take the form of outer products of dimension-specific covariance matrices. We derive some properties of these covariance structures and the corresponding array normal distributions, discuss maximum likelihood and Bayesian estimation of covariance parameters and illustrate the model in an analysis of multivariate longitudinal network data.

Some key words: Gaussian, matrix normal, multiway data, network, tensor, Tucker decomposition.

1 Introduction

This article provides a construction of and estimation for a class of covariance models and Gaussian probability distributions for array data consisting of multi-indexed values $\mathbf{Y} = \{y_{i_1, \dots, i_K} : i_k \in \{1, \dots, m_k\}, k = 1, \dots, K\}$. Such data have become common in many scientific disciplines, including the social and biological sciences. Researchers often gather relational data measured on pairs of units, where the population of units may consist of people, genes, websites or some other set of objects. Data on a single relational variable is often represented by a “sociomatrix”

*Departments of Statistics and Biostatistics , University of Washington, Seattle, Washington 98195-4322. Web: <http://www.stat.washington.edu/hoff/>.

$\mathbf{Y} = \{y_{i,j}, i \in \{1, \dots, m\}, j \in \{1, \dots, m\}, i \neq j\}$, a square matrix with an undefined diagonal, where $y_{i,j}$ represents the relationship between nodes i and j .

Multivariate relational data include multiple relational measurements on the same node set, possibly gathered under different conditions or at different time points. Such data can be represented as a multiway array. For example, in this article we will analyze data on trade of several commodity classes between a set of countries over several years. These data can be represented as a four-way array $\mathbf{Y} = \{y_{i,j,k,t}\}$, where $y_{i,j,k,t}$ records the volume of exports of commodity k from country i to country j in year t . For such data it is often of interest to identify similarities or correlations among data corresponding to the objects of a given index set. For example, one may want to identify nodes of a network that behave similarly across levels of the other factors of the array. For temporal datasets it may be important to describe correlations among data from adjacent time points. In general, it may be desirable to estimate or account for dependencies along each index set of the array.

For matrix valued data, such considerations have led to the use of separable covariance estimates, whereby the covariance of a population of matrices is estimated as being $\text{Cov}[\text{vec}(\mathbf{Y})] = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$. In this parameterization $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ represent covariances among the rows and columns of the matrices, respectively. Such a covariance model may provide a stable and parsimonious alternative to an unrestricted estimate of $\text{Cov}[\text{vec}(\mathbf{Y})]$, the latter being unstable or even unavailable if the dimensions of the sample data matrices are large compared to the sample size. The family of matrix normal distributions with separable covariance matrices is studied in Dawid [1981], and an iterative algorithm for maximum likelihood estimation is given by Dutilleul [1999]. Testing various hypotheses regarding the separability of the covariance structure, or the form of the component matrices, is considered in Lu and Zimmerman [2005], Roy and Khattree [2005], Mitchell et al. [2006] among others. Beyond the matrix-variate case, Galecki [1994] considers a separable covariance model for three-way arrays, but where the component matrices are assumed to have compound symmetry or an autoregressive structure.

In this article we show that the class of separable covariance models for random arrays of arbitrary dimension can be generated with a type of multilinear transformation known as the Tucker product [Tucker, 1964, Kolda, 2006]. Just as a zero-mean multivariate normal vector with a given covariance matrix can be represented as a linear transformation of a vector of independent, standard normal entries, in Section 2 we show that a normal array with separable covariance structure can be represented by a multilinear transformation of an array of independent, standard normal entries. As a result, the calculations involved in obtaining maximum likelihood and Bayesian estimates are made simple via some basic tools of multilinear algebra. In Section 3 we adapt the iterative algorithm of Dutilleul [1999] to the array normal model, and provide a conjugate prior distribution and posterior approximation algorithm for Bayesian inference. Section 4 presents an example data

analysis of trade volume data between pairs of 30 countries in 6 commodity types over 10 years. A discussion of model extensions and directions for further research follows in Section 5.

2 Separable covariance via array-matrix multiplication

2.1 Array notation and basic operations

An array of order K , or K -array, is a map from the product space of K index sets to the real numbers. The different index sets are referred to as the *modes* of the array. The *dimension vector* of an array gives the number of elements in each index set. For example, for a positive integer m_1 , a vector in \mathbb{R}^{m_1} is a one-array with dimension m_1 . A matrix in $\mathbb{R}^{m_1 \times m_2}$ is a two-array with dimension (m_1, m_2) . A K -array \mathbf{Z} with dimension (m_1, \dots, m_K) has elements $\{z_{i_1, \dots, i_K} : i_k \in \{1, \dots, m_k\}, k = 1, \dots, K\}$.

Array *unfolding* refers to the representation of an array by an array of lower order via combinations of various index sets of an array. A useful unfolding is the k -mode matrix unfolding, or k -mode *matricization* [De Lathauwer et al., 2000], in which a K -array \mathbf{Z} is reshaped to form a matrix $\mathbf{Z}_{(k)}$ with m_k rows and $\prod_{j:j \neq k} m_j$ columns. Each column corresponds to the entries of \mathbf{Z} in which the k th index i_k varies from 1 to m_k and the remaining indices are fixed. The assignment of the remaining indices $\{i_j : j \neq k\}$ to columns of $\mathbf{Z}_{(k)}$ is determined by the following ordering on index sets: Letting $\mathbf{i} = (i_1, \dots, i_K)$ and $\mathbf{j} = (j_1, \dots, j_K)$ be two sets of indices, we say $\mathbf{i} < \mathbf{j}$ if $i_k < j_k$ for some k and $i_l \leq j_l$ for all $l > k$. In terms of ordering the columns of the matricization, this means that the index corresponding to a lower-numbered mode “moves faster” than that of a higher-numbered mode.

De Lathauwer et al. [2000] define an array-matrix product via the usual matrix product as applied to matricizations. The k -mode product of an $m_1 \times \dots \times m_K$ array \mathbf{Z} and a $n \times m_k$ matrix \mathbf{A} is obtained by forming the $m_1 \times \dots \times m_{k-1} \times n \times m_{k+1} \times \dots \times m_K$ array from the inversion of the k -mode matricization operation on the matrix $\mathbf{A}\mathbf{Z}_{(k)}$. The resulting array is denoted by $\mathbf{Z} \times_k \mathbf{A}$. Letting \mathbf{F} and \mathbf{G} be matrices of the appropriate sizes, important properties of this product include the following:

- $(\mathbf{Z} \times_j \mathbf{F}) \times_k \mathbf{G} = (\mathbf{Z} \times_k \mathbf{G}) \times_j \mathbf{F} = \mathbf{Z} \times_j \mathbf{F} \times_k \mathbf{G}$
- $(\mathbf{Z} \times_j \mathbf{F}) \times_j \mathbf{G} = \mathbf{Z} \times_j (\mathbf{G}\mathbf{F})$
- $\mathbf{Z} \times_j (\mathbf{F} + \mathbf{G}) = \mathbf{Z} \times_j \mathbf{F} + \mathbf{Z} \times_j \mathbf{G}$.

[De Lathauwer et al., 2000]. A useful extension of the k -mode product is the product of an array \mathbf{Z} with each matrix in a list $\mathbf{A} = \{\mathbf{A}_1, \dots, \mathbf{A}_K\}$ in which $\mathbf{A}_k \in \mathbb{R}^{n_k \times m_k}$, given by

$$\mathbf{Z} \times \mathbf{A} = \mathbf{Z} \times_1 \mathbf{A}_1 \times_2 \dots \times_K \mathbf{A}_K.$$

This has been called the “Tucker operator” or “Tucker product”, [Kolda, 2006], named after the Tucker decomposition for multiway arrays [Tucker, 1964, 1966], and is used for a type of multiway singular value decomposition [De Lathauwer et al., 2000]. A useful calculation involving the Tucker operator is that if $\mathbf{Y} = \mathbf{Z} \times \mathbf{A}$, then

$$\mathbf{Y}_{(k)} = \mathbf{A}_k \mathbf{Z}_{(k)} (\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_{k+1} \otimes \mathbf{A}_{k-1} \otimes \cdots \otimes \mathbf{A}_1).$$

Other properties of the Tucker product can be found in De Lathauwer et al. [2000] and Kolda [2006].

2.2 Separable covariance via the Tucker product

Recall that the general linear group GL_m of nonsingular real matrices \mathbf{A} acts transitively on the space S_m of positive definite $m \times m$ matrices Σ via the transformation $\mathbf{A}\Sigma\mathbf{A}^T$. It is convenient to think of S_m as the set of covariance matrices $\{\text{Cov}[\mathbf{A}\mathbf{z}] : \mathbf{A} \in \text{GL}_m\}$ where \mathbf{z} is an m -variate mean-zero random vector with identity covariance matrix. Additionally, if \mathbf{z} is a vector of independent standard normal random variables, then the distributions of $\mathbf{y} = \mathbf{A}\mathbf{z}$ as \mathbf{A} ranges over GL_m constitute the family of mean-zero vector-valued multivariate normal distributions, which we write as $\mathbf{y} \sim \text{vnorm}(\mathbf{0}, \Sigma)$.

Analogously, let $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2\} \in \text{GL}_{m_1, m_2} \equiv \text{GL}_{m_1} \times \text{GL}_{m_2}$, and let \mathbf{Z} be a $m_1 \times m_2$ random matrix with uncorrelated mean-zero variance-one entries. The covariance structure of the random matrix $\mathbf{Y} = \mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T$ can be described by the $m_1 \times m_1 \times m_2 \times m_2$ covariance array $\text{Cov}[\mathbf{Y}]$ for which the (i_1, i_2, j_1, j_2) entry is equal to $\text{Cov}[y_{i_1, j_1}, y_{i_2, j_2}]$. It is straightforward to show that $\text{Cov}[\mathbf{Y}] = \Sigma_1 \circ \Sigma_2$, where $\Sigma_j = \mathbf{A}_j \mathbf{A}_j^T$, $j = 1, 2$ and “ \circ ” denotes the outer product. This is referred to as a “separable” covariance structure, in which the covariance among elements of \mathbf{Y} can be described by the row covariance Σ_1 and the column covariance Σ_2 . Letting “ $\text{tr}()$ ” be matrix trace and “ \otimes ” the Kronecker product, well-known alternative ways to describe the covariance structure are as follows:

$$\begin{aligned} \text{E}[\mathbf{Y}\mathbf{Y}^T] &= \Sigma_1 \times \text{tr}(\Sigma_2) \\ \text{E}[\mathbf{Y}^T\mathbf{Y}] &= \Sigma_2 \times \text{tr}(\Sigma_1) \\ \text{Cov}[\text{vec}(\mathbf{Y})] &= \Sigma_2 \otimes \Sigma_1. \end{aligned} \tag{1}$$

As $\{\mathbf{A}_1, \mathbf{A}_2\}$ ranges over GL_{m_1, m_2} the covariance array of $\mathbf{Y} = \mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T$ ranges over the space of separable covariance arrays $S_{m_1, m_2} = \{\Sigma_1 \circ \Sigma_2 : \Sigma_1 \in S_{m_1}, \Sigma_2 \in S_{m_2}\}$ [Browne and Shapiro, 1991]. If we additionally assume that the elements of \mathbf{Z} are independent standard normal random variables, then the distributions of $\{\mathbf{Y} = \mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T : \{\mathbf{A}_1, \mathbf{A}_2\} \in \text{GL}_{m_1, m_2}\}$ constitute what are known as the mean-zero matrix normal distributions [Dawid, 1981], which we write as $\mathbf{Y} \sim \text{mnorm}(\mathbf{0}, \Sigma_1 \circ \Sigma_2)$.

Thinking of the matrices \mathbf{Y} and \mathbf{Z} as two-way arrays, the bilinear transformation $\mathbf{Y} = \mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T$ can alternatively be expressed using array-matrix multiplication as $\mathbf{Y} = \mathbf{Z} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 = \mathbf{Z} \times \mathbf{A}$.

Extending this idea further, let \mathbf{Z} be an $m_1 \times \cdots \times m_K$ random array with uncorrelated mean-zero variance-one entries, and define $\text{GL}_{m_1, \dots, m_K}$ to be the set of lists of matrices $\mathbf{A} = \{\mathbf{A}_1, \dots, \mathbf{A}_K\}$ with $\mathbf{A}_k \in \text{GL}_{m_k}$. The Tucker product $\mathbf{Z} \times \mathbf{A}$ induces a transformation on the covariance structure of \mathbf{Z} which shares many features of the analogous bilinear transformation for matrices:

Proposition 2.1 *Let $\mathbf{Y} = \mathbf{Z} \times \mathbf{A}$, where \mathbf{Z} and $\tilde{\mathbf{A}}$ are as above, and let $\Sigma_k = \mathbf{A}_k \mathbf{A}_k^T$. Then*

1. $\text{Cov}[\mathbf{Y}] = \Sigma_1 \circ \cdots \circ \Sigma_K$,
2. $\text{Cov}[\text{vec}(\mathbf{Y})] = \Sigma_K \otimes \cdots \otimes \Sigma_1$,
3. $\text{E}[\mathbf{Y}_{(k)} \mathbf{Y}_{(k)}^T] = \Sigma_k \times \prod_{j:j \neq k} \text{tr}(\Sigma_j)$.

The following result highlights the relationship between array-matrix multiplication and separable covariance structure:

Proposition 2.2 *If $\text{Cov}[\mathbf{Y}] = \Sigma_1 \circ \cdots \circ \Sigma_K$ and $\mathbf{X} = \mathbf{Y} \times_k \mathbf{G}$, then*

$$\text{Cov}[\mathbf{X}] = \Sigma_1 \circ \cdots \circ \Sigma_{k-1} \circ (\mathbf{G} \Sigma_k \mathbf{G}^T) \circ \Sigma_{k+1} \circ \cdots \circ \Sigma_K.$$

This indicates that the class of separable covariance arrays can be obtained by repeated single-mode array-matrix multiplications starting with an array \mathbf{Z} of uncorrelated entries, i.e. for which $\text{Cov}[\mathbf{Z}] = \mathbf{I}_{m_1} \circ \cdots \circ \mathbf{I}_{m_K}$. The class of separable covariance arrays is therefore closed under this group of transformations.

3 Covariance estimation with array normal distributions

3.1 Construction of an array normal class of distributions

Normal probability distributions are useful statistical modeling tools that can represent mean and covariance structure. A family of normal distributions for random arrays with separable covariance structure can be generated as in the vector and matrix cases: Let \mathbf{Z} be an array of independent standard normal entries, and let $\mathbf{Y} = \mathbf{M} + \mathbf{Z} \times \mathbf{A}$ with $\mathbf{M} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ and $\mathbf{A} \in \text{GL}_{m_1, \dots, m_K}$. We say that \mathbf{Y} has an array normal distribution, denoted $\mathbf{Y} \sim \text{anorm}(\mathbf{M}, \Sigma_1 \circ \cdots \circ \Sigma_K)$, where $\Sigma_k = \mathbf{A}_k \mathbf{A}_k^T$.

Proposition 3.1 *The probability density of \mathbf{Y} is given by*

$$p(\mathbf{Y} | \mathbf{M}, \Sigma_1, \dots, \Sigma_K) = (2\pi)^{-m/2} \left(\prod_{k=1}^K |\Sigma_k|^{-m/(2m_k)} \right) \times \exp(-\|(\mathbf{Y} - \mathbf{M}) \times \Sigma^{-1/2}\|^2/2),$$

where $m = \prod_1^K m_k$, $\Sigma^{-1/2} = \{\mathbf{A}_1^{-1}, \dots, \mathbf{A}_K^{-1}\}$ and the array norm $\|\mathbf{Z}\|^2 = \langle \mathbf{Z}, \mathbf{Z} \rangle$ is derived from the inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i_1} \dots \sum_{i_K} x_{i_1, \dots, i_K} y_{i_1, \dots, i_K}$.

Also important for statistical modeling is the idea of replication. If $\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{\text{iid}}{\sim} \text{anorm}(\mathbf{M}, \Sigma_1 \circ \dots \circ \Sigma_K)$, then the array $m_1 \times \dots \times m_K \times n$ array formed by stacking the \mathbf{Y}_i 's together also has an array normal distribution: If $\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{\text{iid}}{\sim} \text{anorm}(\mathbf{M}, \Sigma_1 \circ \dots \circ \Sigma_K)$, then

$$\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_{m_K}) \sim \text{anorm}(\mathbf{M} \circ \mathbf{1}_n, \Sigma_1 \circ \dots \circ \Sigma_K \circ \mathbf{I}_n).$$

This can be shown by computing the joint density of $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ and comparing it to the array normal density.

An important feature of the multivariate normal distribution is that it provides a conditional model of one set of variables given another. Recall, if $\mathbf{y} \sim \text{vnorm}(\boldsymbol{\mu}, \Sigma)$ then the conditional distribution of one subset of elements \mathbf{y}_b of \mathbf{y} given another \mathbf{y}_a is $\text{vnorm}(\boldsymbol{\mu}_{b|a}, \Sigma_{b|a})$, where

$$\begin{aligned} \boldsymbol{\mu}_{b|a} &= \boldsymbol{\mu}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\mathbf{y}_a - \boldsymbol{\mu}_{[a]}) \\ \Sigma_{b|a} &= \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}, \end{aligned}$$

with $\Sigma_{[b,a]}$, for example, being the matrix made up of the entries in the rows of Σ corresponding to b and columns corresponding to a .

A similar result holds for the array normal distribution: Let a and b be non-overlapping subsets of $\{1, \dots, m_1\}$. Let $\mathbf{Y}_b = \{y_{i_1, \dots, i_K} : i_1 \in \mathbf{b}\}$ and $\mathbf{Y}_a = \{y_{i_1, \dots, i_K} : i_1 \in \mathbf{a}\}$ be arrays of dimension $m_{1b} \times m_2 \times \dots \times m_K$ and $m_{1a} \times m_2 \times \dots \times m_K$, where m_{1a} and m_{1b} are the lengths of a and b respectively. The arrays \mathbf{Y}_a and \mathbf{Y}_b are made up of non-overlapping ‘‘slices’’ of the array \mathbf{Y} along the first mode.

Proposition 3.2 *Let $\mathbf{Y} \sim \text{anorm}(\mathbf{M}, \Sigma_1 \circ \dots \circ \Sigma_K)$. The conditional distribution of \mathbf{Y}_b given \mathbf{Y}_a is array normal with mean $\mathbf{M}_{b|a}$ and covariance $\Sigma_{1,b|a} \circ \Sigma_2 \circ \dots \circ \Sigma_K$, where*

$$\begin{aligned} \mathbf{M}_{b|a} &= \mathbf{M}_b + (\mathbf{Y}_a - \mathbf{M}_a) \times_1 (\Sigma_{1[b,a]}(\Sigma_{1[a,a]})^{-1}) \\ \Sigma_{b|a} &= \Sigma_{1[b,b]} - \Sigma_{1[b,a]}(\Sigma_{1[a,a]})^{-1}\Sigma_{1[a,b]}. \end{aligned}$$

Since the conditional distribution is also in the array normal class, successive applications of Proposition 3.2 can be used to obtain the conditional distribution of any subset of the elements of \mathbf{Y} of the form $\{y_{i_1, \dots, i_K} : i_k \in b_k\}$, conditional upon the other elements of the array.

3.2 Estimation for the array normal model

Maximum likelihood estimation: Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n \stackrel{\text{iid}}{\sim} \text{anorm}(\mathbf{M}, \Sigma_1 \circ \dots \circ \Sigma_K)$, or equivalently, $\mathbf{Y} \sim \text{anorm}(\mathbf{M} \circ \mathbf{1}, \Sigma_1 \circ \dots \circ \Sigma_K \circ \mathbf{I}_n)$. For any value of $\Sigma = \{\Sigma_1, \dots, \Sigma_K\}$, the value of \mathbf{M} that

maximizes $p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma})$ is the value that minimizes the residual mean squared error:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|(\mathbf{Y}_i - \mathbf{M}) \times \mathbf{\Sigma}^{-1/2}\|^2 &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{Y}_i - \mathbf{M}, (\mathbf{Y}_i - \mathbf{M}) \times \mathbf{\Sigma}^{-1} \rangle \\
&= \langle \mathbf{M}, \mathbf{M} \times \mathbf{\Sigma}^{-1} \rangle - 2\langle \mathbf{M}, \bar{\mathbf{Y}} \times \mathbf{\Sigma}^{-1} \rangle + c_1(\mathbf{Y}, \mathbf{\Sigma}) \\
&= \langle \mathbf{M} - \bar{\mathbf{Y}}, (\mathbf{M} - \bar{\mathbf{Y}}) \times \mathbf{\Sigma}^{-1} \rangle + c_2(\mathbf{Y}, \mathbf{\Sigma}) \\
&= \|(\mathbf{M} - \bar{\mathbf{Y}}) \times \mathbf{\Sigma}^{-1/2}\|^2 + c_2(\mathbf{Y}, \mathbf{\Sigma}).
\end{aligned}$$

This is uniquely minimized in \mathbf{M} by $\bar{\mathbf{Y}} = \sum \mathbf{Y}_i/n$, and so $\hat{\mathbf{M}} = \bar{\mathbf{Y}}$ is the MLE of \mathbf{M} . The MLE of $\mathbf{\Sigma}$ does not have a closed form expression. However, it is possible to maximize $p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma})$ in $\mathbf{\Sigma}_k$, given values of the other covariance matrices. Letting $\mathbf{E} = \mathbf{Y} - \mathbf{M} \circ \mathbf{1}_n$, the likelihood as a function of $\mathbf{\Sigma}_k$ can be expressed as $p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}_k|^{-nm/(2m_k)} \exp\{-\|\mathbf{E} \times \{\mathbf{\Sigma}_1^{-1/2}, \dots, \mathbf{\Sigma}_K^{-1/2}, \mathbf{I}_n\}\|^2/2\}$. Since for any array \mathbf{Z} and mode k we have $\|\mathbf{Z}\|^2 = \|\mathbf{Z}_{(k)}\|^2 = \text{tr}(\mathbf{Z}_{(k)}\mathbf{Z}_{(k)}^T)$, the norm in the likelihood can be written as

$$\begin{aligned}
\|\mathbf{E} \times \mathbf{\Sigma}^{-1/2}\|^2 &= \|\tilde{\mathbf{E}} \times_k \mathbf{\Sigma}_k^{-1/2}\|^2 \\
&= \text{tr}(\mathbf{\Sigma}_k^{-1/2} \tilde{\mathbf{E}}_{(k)} \tilde{\mathbf{E}}_{(k)}^T \mathbf{\Sigma}_k^{-1/2}) \\
&= \text{tr}(\mathbf{\Sigma}_k^{-1} \tilde{\mathbf{E}}_{(k)} \tilde{\mathbf{E}}_{(k)}^T),
\end{aligned}$$

where $\tilde{\mathbf{E}} = \mathbf{E} \times \{\mathbf{\Sigma}_1^{-1/2}, \dots, \mathbf{\Sigma}_{k-1}^{-1/2}, \mathbf{I}_k, \mathbf{\Sigma}_{k+1}^{-1/2}, \dots, \mathbf{\Sigma}_K^{-1/2}, \mathbf{I}_n\}$ is the residual array standardized along each dimension except k . Writing $\mathbf{S}_k = \tilde{\mathbf{E}}_{(k)} \tilde{\mathbf{E}}_{(k)}^T$, we have

$$p(\mathbf{Y}|\mathbf{M}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}_k|^{-nm/(2m_k)} \text{etr}\{-\mathbf{\Sigma}_k^{-1} \mathbf{S}_k/2\}$$

as a function of $\mathbf{\Sigma}_k$, and so if \mathbf{S}_k is of full rank then the unique maximizer in $\mathbf{\Sigma}_k$ is given by $\hat{\mathbf{\Sigma}}_k = \mathbf{S}_k/n_k$, where $n_k = nm/m_k = n \times \prod_{j \neq k} m_j$ is the number of columns of $\mathbf{Y}_{(k)}$, i.e. the “sample size” for the k th mode. This suggests the following iterative algorithm for obtaining the MLE of $\mathbf{\Sigma}$: Letting $\mathbf{E} = \mathbf{Y} - \bar{\mathbf{Y}} \circ \mathbf{1}_n$ and given an initial value of $\mathbf{\Sigma}$, for each $k \in \{1, \dots, K\}$

1. compute $\tilde{\mathbf{E}} = \mathbf{E} \times \{\mathbf{\Sigma}_1^{-1/2}, \dots, \mathbf{\Sigma}_{k-1}^{-1/2}, \mathbf{I}_k, \mathbf{\Sigma}_{k+1}^{-1/2}, \dots, \mathbf{\Sigma}_K^{-1/2}, \mathbf{I}_n\}$ and $\mathbf{S}_k = \tilde{\mathbf{E}}_{(k)} \tilde{\mathbf{E}}_{(k)}^T$;
2. set $\mathbf{\Sigma}_k = \mathbf{S}_k/n_k$, where $n_k = n \times \prod_{j \neq k} m_j$.

Each iteration increases the likelihood, and so the procedure can be seen as a type of block coordinate descent algorithm [Tseng, 2001]. For the matrix normal case, this algorithm was proposed by Dutilleul [1999] and is sometimes called the “flip-flop” algorithm. Note that the scales of $\{\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_K\}$ are not separately identifiable from the likelihood: Replacing $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ with $c\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2/c$ yield the same probability distribution for \mathbf{Y} , and so the scales of the MLEs will depend on the initial values.

Bayesian estimation: Estimation of high-dimensional parameters often benefits from a complexity penalty, e.g. a penalty on the magnitude of the parameters. Such penalties can often be expressed as prior distributions, and so penalized likelihood estimation can be done in the context of Bayesian inference. With this in mind, we consider semiconjugate prior distributions for the array normal model, and their associated posterior distributions.

A conjugate prior distribution for the for the multivariate normal model $\mathbf{y}_1 \dots \mathbf{y}_n \stackrel{\text{iid}}{\sim} \text{vnorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma})$, where $p(\boldsymbol{\Sigma})$ is an inverse-Wishart density and $p(\boldsymbol{\mu}|\boldsymbol{\Sigma})$ is multivariate normal density with prior mean $\boldsymbol{\mu}_0$ and prior (conditional) covariance $\boldsymbol{\Sigma}/\kappa_0$. The parameter κ_0 can be thought of as a “prior sample size,” as the the prior covariance $\boldsymbol{\Sigma}/\kappa_0$ for $\boldsymbol{\mu}$ is the same as that of a sample average based on κ_0 observations. Under this prior distribution, the conditional distribution of $\boldsymbol{\mu}$ given the data and $\boldsymbol{\Sigma}$ is multivariate normal, and the condition distribution of $\boldsymbol{\Sigma}$ given the data is inverse-Wishart. An analogous result holds for the array normal model: If

$$\begin{aligned}\mathbf{M}|\boldsymbol{\Sigma} &\sim \text{anorm}(\mathbf{M}_0, \boldsymbol{\Sigma}_1 \circ \dots \circ \boldsymbol{\Sigma}_K/\kappa_0) \\ \boldsymbol{\Sigma}_k &\sim \text{inverse-Wishart}(\mathbf{S}_{0k}^{-1}, \nu_{0k})\end{aligned}$$

and $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K$ are independent, then straightforward calculations show that

$$\begin{aligned}\mathbf{M}|\mathbf{Y}, \boldsymbol{\Sigma} &\sim \text{anorm}([\kappa_0\mathbf{M}_0 + n\bar{\mathbf{Y}}]/[\kappa_0 + n], \boldsymbol{\Sigma}_1 \circ \dots \circ \boldsymbol{\Sigma}_K/[\kappa_0 + n]) \\ \boldsymbol{\Sigma}_k|\mathbf{Y}, \boldsymbol{\Sigma}_{-k} &\sim \text{inverse-Wishart}([\mathbf{S}_{0k} + \mathbf{S}_k + \mathbf{R}_{(k)}\mathbf{R}_{(k)}^T]^{-1}, \nu_{0k} + n_k),\end{aligned}$$

where \mathbf{S}_k and n_k are as in the coordinate descent algorithm for maximum likelihood estimation, and $\mathbf{R} = \sqrt{\frac{\kappa_0 n}{\kappa_0 + n}}(\bar{\mathbf{Y}} - \mathbf{M}_0) \times \{\boldsymbol{\Sigma}_1^{-1/2}, \dots, \boldsymbol{\Sigma}_{k-1}^{-1/2}, \mathbf{I}, \boldsymbol{\Sigma}_{k+1}^{-1/2}, \dots, \boldsymbol{\Sigma}_K^{-1/2}\}$.

As noted above, the scales of $\{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$ are not separately identifiable from the likelihood. This makes the prior and posterior distributions of the scales of the $\boldsymbol{\Sigma}_k$ ’s difficult to specify or interpret. As a remedy, we consider reparameterizing the prior distribution for $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K$ to include a parameter representing the total variance in the data. Parameterizing $\mathbf{S}_{0k} = \gamma \boldsymbol{\Sigma}_{0k}$ for each $k \in \{1, \dots, K\}$, the prior expected total variation of \mathbf{Y}_i , $\text{tr}(\text{Cov}[\text{vec}(\mathbf{Y}_i)])$, is

$$\begin{aligned}\mathbb{E}[\text{tr}(\text{Cov}[\text{vec}(\mathbf{Y})])] &= \mathbb{E}[\text{tr}(\boldsymbol{\Sigma}_K \otimes \dots \otimes \boldsymbol{\Sigma}_1)] \\ &= \mathbb{E}\left[\prod_{k=1}^K \text{tr}(\boldsymbol{\Sigma}_k)\right] \\ &= \prod_{k=1}^K \text{tr}(\mathbb{E}[\boldsymbol{\Sigma}_k]) = \gamma^K \prod_{k=1}^K \text{tr}(\boldsymbol{\Sigma}_{0k})/(\nu_{0k} - m_k - 1).\end{aligned}$$

A simple default choice for $\boldsymbol{\Sigma}_{0k}$ and ν_{0k} would be $\boldsymbol{\Sigma}_{0k} = \mathbf{I}_{m_k}/m_k$ and $\nu_{0,k} = m_k + 2$, for which $\mathbb{E}[\text{tr}(\boldsymbol{\Sigma}_k)] = \gamma$ and the expected value for the total variation is γ^K . Given prior expectations about the total variance, the value of γ could be set accordingly. Alternatively, a prior distribution could

be placed on γ : If $\gamma \sim \text{gamma}(a, b)$ with prior mean a/b , then conditional on $\Sigma_1, \dots, \Sigma_K$, we have

$$\gamma | \Sigma_1, \dots, \Sigma_K \sim \text{gamma} \left(a + \sum \nu_{0k} m_k / 2, b + \sum \text{tr}(\Sigma_k^{-1} \Sigma_{0k}) / 2 \right).$$

The full conditional distributions of $\{\mathbf{M}, \Sigma_1, \dots, \Sigma_K, \gamma\}$ can be used to implement a Gibbs sampler, in which each parameter is sampled in turn from its full conditional distribution, given the current values of the other parameters. This algorithm generates a Markov chain having a stationary density equal to $p(\mathbf{M}, \Sigma_1, \dots, \Sigma_K, \gamma | \mathbf{Y})$, samples from which can be used to approximate posterior quantities of interest. Such an algorithm is implemented in the data analysis example in the next section.

4 Example: International trade

The United Nations gathers yearly trade data between countries of the world and disseminates this information at the UN Comtrade website <http://comtrade.un.org/>. In this section we analyze trade among pairs of countries over several years and in several different commodity categories. Specifically, the data take the form of a four-mode array $\mathbf{Y} = \{y_{i,j,k,t}\}$ where

- $i \in \{1, \dots, 30 = m\}$ indexes the exporting nation;
- $j \in \{1, \dots, 30 = m\}$ indexes the importing nation;
- $k \in \{1, \dots, 6 = p\}$ indexes the commodity type;
- $t \in \{1, \dots, 10 = n\}$ indexes the year.

The thirty countries were selected to make the data as complete as possible, resulting in a set of mostly large or developed countries with high gross domestic products and trade volume. The six commodity types include (1) chemicals, (2) inedible crude materials not including fuel, (3) food and live animals, (4) machinery and transport equipment, (5) textiles and (6) manufactured goods. The years represented in the dataset include 1996 through 2005. As trade between countries is relatively stable across years, we analyze the yearly change in log trade values, measured in 2000 US dollars. For example, $y_{1,2,1,1}$ is the log-dollar increase in the value of chemicals exported from Australia to Austria from 1995 to 1996. We note that exports of a country to itself are not defined, so $y_{i,i,j,t}$ is “not available” and can be treated as missing at random.

We model these data as $\mathbf{Y} = \mathbf{M} \circ \mathbf{1}_n + \mathbf{E}$, where \mathbf{M} is an $m \times m \times p$ array of means specific to exporter-importer-commodity combinations, and \mathbf{E} is an $m \times m \times p \times n$ array of residuals. Of interest here is how the deviations \mathbf{E} of the data from the mean may be correlated across exporters, importers and commodities. One possible model for this residual variation would be to treat the p -dimensional residual vectors corresponding to each of the $m \times (m - 1) \times n = 8700$ exporter-importer-year

combinations as independent samples from a p -variate multivariate normal distribution. However, to accommodate potential temporal correlation (beyond that already accounted for by taking \mathbf{Y} to be the lagged log trade values), the $p \times n$ residual matrices corresponding to each of the $m \times (m - 1) = 870$ exporter-importer pairs could be modeled as independent samples from a matrix normal distribution, with two separate covariance matrices representing commodity and temporal correlation. This latter model can be described by an array normal model as

$$\mathbf{Y} \sim \text{anorm}(\mathbf{M} \circ \mathbf{1}_n, \mathbf{I}_m \circ \mathbf{I}_m \circ \Sigma_3 \circ \Sigma_4), \quad (2)$$

where Σ_3 and Σ_4 describe covariance among commodities and time points, respectively. However, it is natural to consider the possibility that there will be correlation of residuals attributable to exporters and importers. For example, countries with similar economies may exhibit correlations in their trade patterns. With this in mind, we will also fit the following model:

$$\mathbf{Y} \sim \text{anorm}(\mathbf{M} \circ \mathbf{1}_n, \Sigma_1 \circ \Sigma_2 \circ \Sigma_3 \circ \Sigma_4). \quad (3)$$

We consider Bayesian analysis for both of these models based on the prior distributions described at the end of the last section. The prior distribution for each Σ_k matrix being estimated is given by $\Sigma_k \sim \text{inverse-Wishart}(m_k \mathbf{I}_{m_k} / \gamma, m_k + 2)$, with the hyperparameter γ set so that $\gamma^K = \|\mathbf{Y} - \bar{\mathbf{Y}} \circ \mathbf{1}_n\|^2$. As described in the previous section, this weakly centers the total variation of \mathbf{Y} under the model around the empirically observed value, similar to an empirical Bayes approach or unit information prior distribution [Kass and Wasserman, 1995]. The prior distribution for \mathbf{M} conditional on Σ is $\mathbf{M} \sim \text{anorm}(\mathbf{0}, \Sigma_1 \circ \Sigma_2 \circ \Sigma_3)$, where $\Sigma_1 = \Sigma_2 = \mathbf{I}_m$ for model 2.

Posterior distributions of parameters for both model 2 and 3 can be obtained using the results of the previous section with minor modifications. Under both models the $m \times m \times p$ arrays $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ corresponding to the $n = 10$ time-points are not independent, but correlated according to Σ_4 . The full conditional distribution of \mathbf{M} is still given by an array normal distribution, but the mean and variance are now as follows:

$$\begin{aligned} \mathbb{E}[\mathbf{M} | \mathbf{Y}, \Sigma] &= \frac{\kappa_0 \mathbf{M}_0 + \sum_{i=1}^n c_i \tilde{\mathbf{Y}}_i}{\kappa_0 + \sum c_i^2} \\ \text{Var}[\mathbf{M} | \mathbf{Y}, \Sigma] &= \Sigma_1 \circ \Sigma_2 \circ \Sigma_3 \circ \Sigma_4 / (\kappa_0 + \sum c_i^2) \end{aligned}$$

where $\tilde{\mathbf{Y}}_1, \dots, \tilde{\mathbf{Y}}_n$ are the $m \times m \times p$ arrays obtained from the first three modes of the transformed array $\tilde{\mathbf{Y}} = \mathbf{Y} \times_4 \Sigma_4^{-1/2}$, and c_1, \dots, c_n are the elements of the vector $\mathbf{c} = \Sigma_4^{-1/2} \mathbf{1}$. Additionally, the time dependence makes it difficult to integrate $p(\mathbf{M}, \Sigma | \mathbf{Y})$ as was possible in the independent case. As a result, we use a Gibbs sampler that proceeds by sampling Σ_k from its full conditional distribution $p(\Sigma_k | \mathbf{Y}, \mathbf{M}, \Sigma_{-k})$ as opposed to the $p(\Sigma_k | \mathbf{Y}, \Sigma_{-k})$ as before. This full conditional distribution is still a member of the inverse-Wishart family:

$$\Sigma_k | \mathbf{Y}, \mathbf{M}, \Sigma_{-k} \sim \text{inverse-Wishart}([\mathbf{S}_{0k} + \mathbf{E}_{(k)} \mathbf{E}_{(k)}^T + \mathbf{R}_{(k)} \mathbf{R}_{(k)}^T]^{-1}, \nu_{0k} + n_k \times [1 + 1/n]),$$

where $\mathbf{E}_{(1)}$, for example, is the k -mode matricization of $(\mathbf{Y} - \mathbf{M} \circ \mathbf{1}_n) \times \{\mathbf{I}_m, \boldsymbol{\Sigma}_2^{-1/2}, \boldsymbol{\Sigma}_3^{-1/2}, \boldsymbol{\Sigma}_4^{-1/2}\}$ and $\mathbf{R}_{(1)}$ is the k -mode matricization of $(\mathbf{M} - \mathbf{M}_0) \times \{\mathbf{I}, \boldsymbol{\Sigma}_2^{-1/2}, \boldsymbol{\Sigma}_3^{-1/2}\}$.

Separate Markov chains for each of the two models were generated using 205,000 iterations of the Gibbs sampler discussed above. The first 5,000 iterations were dropped from each chain to allow for convergence to the stationary distribution, and parameter values were saved every 40th iteration thereafter, resulting in 5,000 parameter values with which to approximate the posterior distributions. Mixing of the Markov chain was assessed by computing the “effective sample size”, or equivalent number of independent simulations, of several summary parameters. For the full model, effective sample sizes of $\gamma_0 = \text{tr}(\boldsymbol{\Sigma}_1 \otimes \cdots \otimes \boldsymbol{\Sigma}_4)$, $\gamma_1 = \text{tr}(\boldsymbol{\Sigma}_1)$, \dots , $\gamma_4 = \text{tr}(\boldsymbol{\Sigma}_4)$ were computed to be 2,545, 904, 960, 548 and 1,734. Note that $\gamma_1, \dots, \gamma_4$ are not separately identifiable from the data, resulting in poorer mixing than γ_0 , which is identifiable. For the reduced model, the effective sample sizes of γ_0 , γ_3 and γ_4 were 4,281, 1,194 and 1,136.

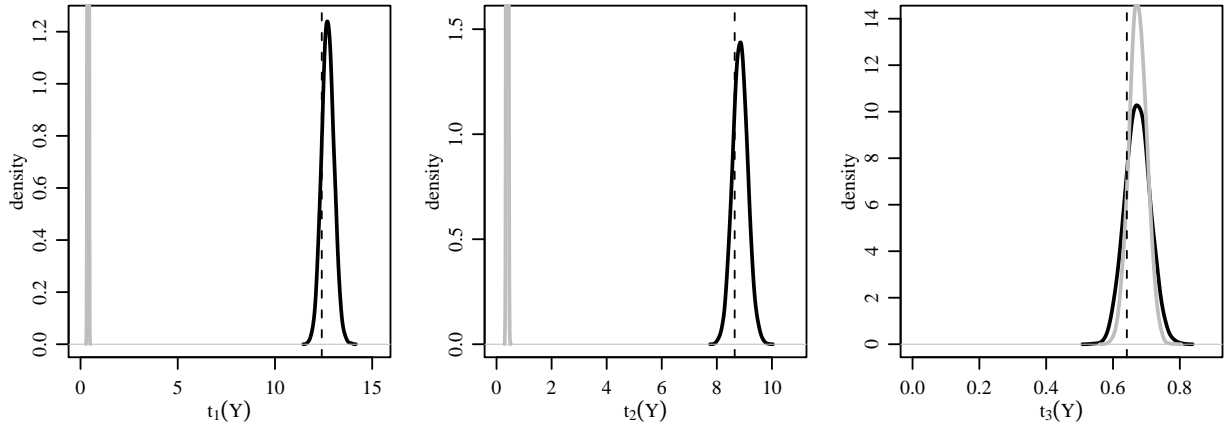


Figure 1: Posterior predictive distributions for summary statistics. The gray density represents is under the reduced model, the black under the full. The vertical dashed line is the observed value of the statistic.

The fits of the two models can be compared using posterior predictive evaluations [Rubin, 1984]: To evaluate the fit of a model, the observed value of a summary statistic $t(\mathbf{Y})$ can be compared to values $t(\tilde{\mathbf{Y}})$ for which $\tilde{\mathbf{Y}}$ is simulated from the posterior predictive distribution. A discrepancy between $t(\mathbf{Y})$ and the distribution of $t(\tilde{\mathbf{Y}})$ indicates that the model is not capturing the aspect of the data represented by $t(\cdot)$. For illustration, we use such checks here to evaluate evidence that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not equal to the identity, or equivalently, that model 1 exhibits lack of fit as compared to model 2. To obtain a summary statistic evaluating evidence of a non-identity covariance matrix for $\boldsymbol{\Sigma}_1$, we first subtract the sample mean from the data array to obtain $\mathbf{E} = \mathbf{Y} - \bar{\mathbf{Y}}$, and then compute $\mathbf{S}_1 = (\mathbf{E}_{(1)}\mathbf{E}_{(1)}^T)$. The $m \times m$ matrix \mathbf{S}_1 is a sample measure of covariance among exporting countries. We then obtain a scaled version $\tilde{\mathbf{S}}_1 = \mathbf{S}_1/\text{tr}(\mathbf{S}_1)$, and compare it to a scaled version of

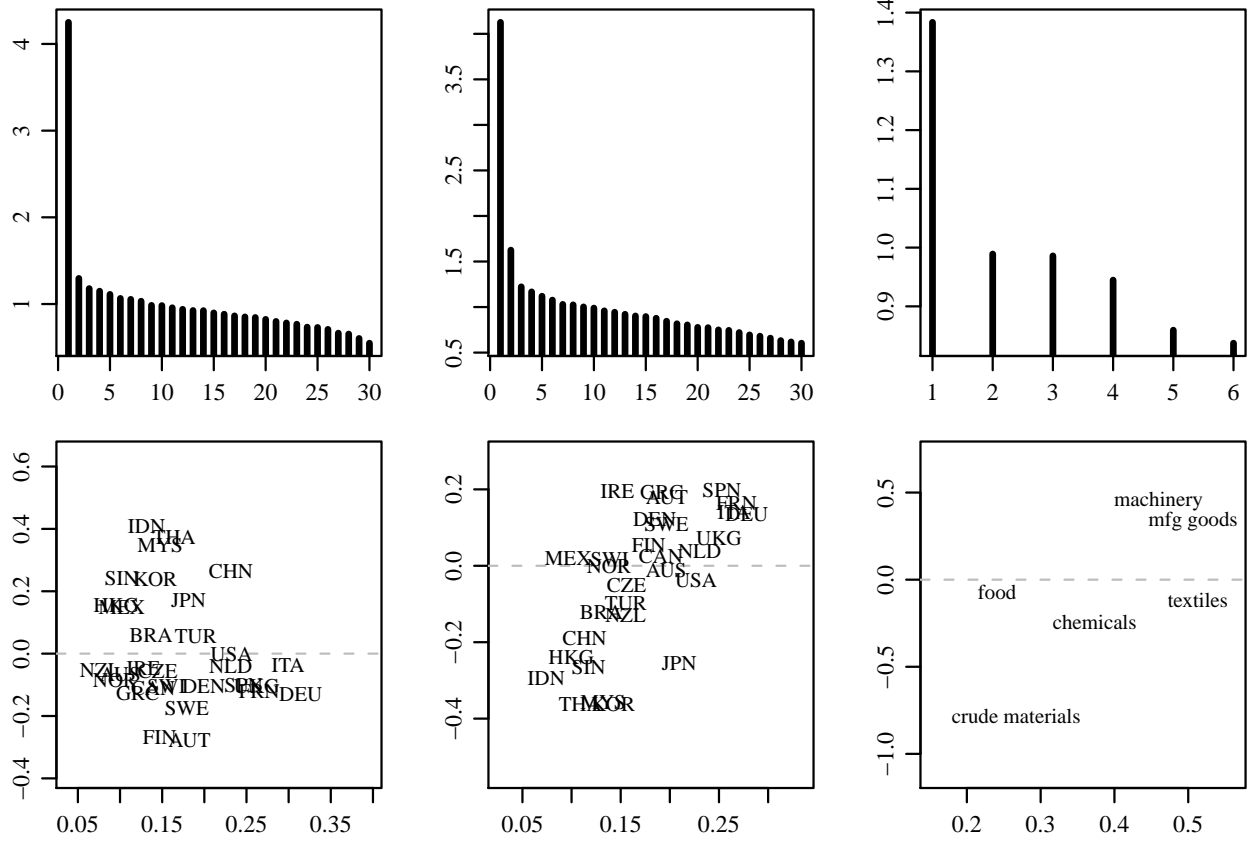


Figure 2: Estimates of correlation matrices corresponding to Σ_1 , Σ_2 and Σ_3 . The first panel in each row plots the eigenvalues of each correlation matrix, the second plots the first two eigenvectors.

the identity matrix:

$$t_1(\mathbf{Y}) = \log |\tilde{\mathbf{S}}_1| - \log |\mathbf{I}/m| = \log |\tilde{\mathbf{S}}_1| + m \log m.$$

Note that the minimum value of this statistic occurs when $\tilde{\mathbf{S}}_1 = \mathbf{I}/m$, and so in some sense it provides a simple scalar measure of how the sample covariance among exporters differs from a scaled identity matrix. Similarly, we construct $t_2(\mathbf{Y})$ and $t_3(\mathbf{Y})$ measuring sample covariance along the second and third modes of the data array. We include t_3 to contrast with t_1 and t_2 , as both the full and reduced models include covariance parameters for the third dimension of the array.

Figure 1 plots posterior predictive densities for $t_1(\tilde{\mathbf{Y}})$, $t_2(\tilde{\mathbf{Y}})$ and $t_3(\tilde{\mathbf{Y}})$ under both the full and reduced models, and compares these densities to the observed values of the statistics. The reduced model exhibits substantial lack of fit in terms of its inability to predict datasets that resemble the observed in terms of t_1 and t_2 . In other words, a model that assumes i.i.d. structure along the first two modes of the array does not fit the data. In terms of covariance among commodities along the third mode, neither model exhibits substantial lack of fit as measured by t_3 .

Figure 2 describes posterior mean estimates of the correlation matrices corresponding to Σ_1 , Σ_2 and Σ_3 . The two panels in each column plot the eigenvalues and the first two eigenvectors each of the three correlation matrices. The eigenvalues for all three suggest the possibility of modeling the covariances matrices with factor analytic structure, i.e. letting $\Sigma_k = \mathbf{A}\mathbf{A}^T + \text{diag}(b_1^2, \dots, b_{m_k}^2)$, where \mathbf{A} is an $m_k \times r$ matrix with $r < m_k$. This possibility is described further in the Discussion. The second row of Figure 2 describes correlations among exporters, importers and commodities. The first two plots show that much of the correlation among exporters and among importers is related to geography, with countries having similar eigenvector values typically being near each other geographically. The third plot in the row indicates correlation among commodities of a similar type: Moving up and to the right from “crude materials,” the commodities are essentially in order of the extent to which they are finished goods.

5 Discussion

This article has proposed a class of array normal distributions with separable covariance structure for the modeling of array data, particularly multivariate relational arrays. It has been shown that the class of separable covariance models can be generated by a type of array-matrix multiplication known as the Tucker product. Maximum likelihood estimation and Bayesian inference in Gaussian versions of such models are feasible using relatively simple properties of array-matrix multiplication. These types of models can be useful for describing covariance within the index sets of an array dataset. In an example involving longitudinal trade data, we have shown that a full array normal model provides a better fit than a matrix normal model, which includes a covariance matrix for only two of the four modes of the data array.

One open area of research is finding conditions for the existence and uniqueness of the MLE for the array normal model. In fact, such conditions for even the simple matrix normal case are not fully established. Each step of the iterative MLE algorithm of Dutilleul [1999] is well defined as long as $n \geq \max\{m_1/m_2, m_2/m_1\} + 1$, suggesting that this condition might be sufficient to provide existence. Unfortunately, it is easy to find examples where this condition is met but the likelihood is unbounded, and no MLE exists. In contrast, Srivastava et al. [2008] show the MLE exists and is essentially unique if $n > \max\{m_1, m_2\}$, a much stronger requirement. However, they do not show that this condition is necessary. Unpublished results of this author suggest that this latter condition can be relaxed, and that the existence of the MLE depends on the minimal possible rank of $\sum_{i=1}^n \mathbf{Y}_i \mathbf{U}_k \mathbf{U}_k^T \mathbf{Y}_i^T$ for $k \leq \min\{m_1, m_2\}$, where \mathbf{U}_k is an $m_2 \times k$ matrix with orthonormal columns. However, such conditions are somewhat opaque, and it is not clear that they could easily be generalized to the array normal case.

A potentially useful model variation would be to consider imposing simplifying structure on the

component matrices. For example, a normal factor analysis model for a random vector $\mathbf{y} \in \mathbb{R}^p$ posits that $\mathbf{y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{z} + \mathbf{D}\boldsymbol{\epsilon}$ where $\mathbf{z} \in \mathbb{R}^r$, $r < p$ and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ are uncorrelated standard normal vectors and \mathbf{D} is a diagonal matrix. The resulting covariance matrix is given by $\text{Cov}[\mathbf{y}] = \mathbf{B}\mathbf{B}^T + \mathbf{D}^2$, in which the “interesting” part $\mathbf{B}\mathbf{B}^T$ is of rank r . The natural extension to random arrays is $\mathbf{Y}_i = \mathbf{M} + \mathbf{Z} \times \{\mathbf{B}_1, \dots, \mathbf{B}_K\} + \mathbf{E} \times \{\mathbf{D}_1, \dots, \mathbf{D}_K\}$ where $\mathbf{Z} \in \mathbb{R}^{r_1 \times \dots \times r_K}$ and $\mathbf{E} \in \mathbb{R}^{p_1 \times \dots \times p_K}$ are uncorrelated standard normal arrays. This induces the covariance matrix $\text{Cov}[\text{vec}(\mathbf{Y})] = (\mathbf{B}_K \mathbf{B}_K^T) \otimes \dots \otimes (\mathbf{B}_1 \mathbf{B}_1^T) + \mathbf{D}_K^2 \otimes \dots \otimes \mathbf{D}_1^2$. This is essentially the model-based analogue of the higher-order SVD of De Lathauwer et al. [2000], in the same way that the usual factor analysis model for vector-valued data is analogous to the matrix SVD. Alternatively, in some cases it may be desirable to fit a factor-analytic structure for the covariances of some modes of the array while estimating others as unstructured. This can be achieved with a model of the following form:

$$\mathbf{Y} = \mathbf{M} + \mathbf{Z} \times \{\boldsymbol{\Sigma}_1^{1/2}, \dots, \boldsymbol{\Sigma}_k^{1/2}, \mathbf{B}_{k+1}, \dots, \mathbf{B}_K\} + \mathbf{E} \times \{\boldsymbol{\Sigma}_1^{1/2}, \dots, \boldsymbol{\Sigma}_k^{1/2}, \mathbf{D}_{k+1}, \dots, \mathbf{D}_K\}$$

where $\mathbf{Z} \in \mathbb{R}^{p_1 \times \dots \times p_k \times r_{k+1} \times \dots \times r_K}$, and $\mathbf{E} \in \mathbb{R}^{p_1 \times \dots \times p_K}$. The resulting covariance for \mathbf{Y} is given by

$$\text{Cov}[\text{vec}(\mathbf{Y})] = (\mathbf{B}_K \mathbf{B}_K^T + \mathbf{D}_K^2) \otimes \dots \otimes (\mathbf{B}_{k+1} \mathbf{B}_{k+1}^T + \mathbf{D}_{k+1}^2)^T \otimes \boldsymbol{\Sigma}_k \otimes \dots \otimes \boldsymbol{\Sigma}_1,$$

which is separable, and so is in the array normal class. Such a factor model may be useful if some modes of the array have very high dimensions, and rank-reduced estimates of the corresponding covariance matrices are desired.

An additional model extension would be to accommodate non-normal continuous or discrete array data, for example, dynamic binary networks. This can be done by embedding the array normal model within a generalized linear model, or within an ordered probit model for ordinal response data. For example, if \mathbf{Y} is a three-way binary array, an array normal probit model would posit a latent array $\mathbf{Z} \sim \text{anorm}(\mathbf{M}, \boldsymbol{\Sigma}_1 \circ \boldsymbol{\Sigma}_2 \circ \boldsymbol{\Sigma}_3)$ which determines \mathbf{Y} via $y_{i,j,k} = \delta_{(0,\infty)}(z_{i,j,k})$.

Computer code and data for the example in Section 5 is available at my website: <http://www.stat.washington.edu/~hoff>.

Appendix

PROOF OF PROPOSITION 3.1. Let $\mathbf{Y} = \mathbf{Z} \times \mathbf{A}$ where the elements of \mathbf{Z} are uncorrelated, have expectation zero and variance one. Using the fact that $\mathbf{Y}_{(1)} = \mathbf{A}_1 \mathbf{Z}_{(1)} \mathbf{B}^T$ where $\mathbf{B} = (\mathbf{A}_K \otimes \dots \otimes \mathbf{A}_2)$ [Kolda, 2006, Proposition 4.3], we have

$$\begin{aligned} \text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{Y}_{(1)}) &= \text{vec}(\mathbf{A}_1 \mathbf{Z}_{(1)} \mathbf{B}^T) \\ &= (\mathbf{B} \otimes \mathbf{A}_1) \text{vec}(\mathbf{Z}_{(1)}) = (\mathbf{B} \otimes \mathbf{A}_1) \text{vec}(\mathbf{Z}). \end{aligned}$$

The covariance of $\text{vec}(\mathbf{Y})$ is then

$$\begin{aligned}
\mathbb{E}[\text{vec}(\mathbf{Y})\text{vec}(\mathbf{Y})^T] &= (\mathbf{B} \otimes \mathbf{A}_1)\mathbb{E}[\text{vec}(\mathbf{Z})\text{vec}(\mathbf{Z})^T](\mathbf{B} \otimes \mathbf{A}_1)^T \\
&= (\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1)\mathbf{I}(\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1)^T \\
&= (\mathbf{A}_K\mathbf{A}_K^T) \otimes \cdots \otimes (\mathbf{A}_1\mathbf{A}_1^T) \\
&= \boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1,
\end{aligned}$$

where $\boldsymbol{\Sigma}_k = \mathbf{A}_k\mathbf{A}_k^T$. This proves the second statement in the proposition. The first statement follows from how the “vec” operation is applied to arrays. For the third statement, consider the calculation of $\mathbb{E}[\mathbf{Y}_{(1)}\mathbf{Y}_{(1)}^T]$, again using the fact that $\mathbf{Y}_{(1)} = \mathbf{A}_1\mathbf{Z}_{(1)}\mathbf{B}^T$:

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}_{(1)}\mathbf{Y}_{(1)}^T] &= \mathbf{A}_1\mathbb{E}[\mathbf{Z}_{(1)}\mathbf{B}^T\mathbf{B}\mathbf{Z}_{(1)}^T]\mathbf{A}_1^T \\
&= \mathbf{A}_1\mathbb{E}[\mathbf{X}\mathbf{X}^T]\mathbf{A}_1^T,
\end{aligned} \tag{4}$$

where $\mathbf{X} = \mathbf{Z}_{(1)}\mathbf{B}^T$. Because the elements of \mathbf{Z} are all independent, mean zero and variance one, the rows of \mathbf{X} are independent with mean zero and variance $\mathbf{B}\mathbf{B}^T$. Thus $\mathbb{E}[\mathbf{X}\mathbf{X}^T] = \text{tr}(\mathbf{B}\mathbf{B}^T)\mathbf{I}$. Combining this with (4) gives

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}_{(1)}\mathbf{Y}_{(1)}^T] &= \mathbf{A}_1\mathbf{A}_1^T\text{tr}(\mathbf{B}\mathbf{B}^T) \\
&= \mathbf{A}_1\mathbf{A}_1^T\text{tr}([\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_2][\mathbf{A}_K^T \otimes \cdots \otimes \mathbf{A}_2^T]) \\
&= \boldsymbol{\Sigma}_1\text{tr}(\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1) \\
&= \boldsymbol{\Sigma}_1 \prod_{k=2}^K \text{tr}(\boldsymbol{\Sigma}_k).
\end{aligned}$$

Calculation of $\mathbb{E}[\mathbf{Y}_{(k)}\mathbf{Y}_{(k)}^T]$ for other values of k is similar. \square

PROOF OF PROPOSITION 3.2. We calculate $\mathbb{E}[\text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T]$ for the case that $\mathbf{X} = \mathbf{Y} \times_1 \mathbf{G}$:

$$\begin{aligned}
\text{vec}(\mathbf{X}) = \text{vec}(\mathbf{X}_{(1)}) &= \text{vec}(\mathbf{G}\mathbf{Y}_{(1)}) \\
&= \text{vec}(\mathbf{G}\mathbf{Y}_{(1)}\mathbf{I}) \\
&= (\mathbf{I} \otimes \mathbf{G})\text{vec}(\mathbf{Y}), \text{ so} \\
\mathbb{E}[\text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T] &= (\mathbf{I} \otimes \mathbf{G})\mathbb{E}[\text{vec}(\mathbf{Y})\text{vec}(\mathbf{Y})^T](\mathbf{I} \otimes \mathbf{G})^T \\
&= (\mathbf{I} \otimes \mathbf{G})(\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_1)(\mathbf{I} \otimes \mathbf{G}^T) \\
&= [(\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_2) \otimes (\mathbf{G}\boldsymbol{\Sigma}_1)] [\mathbf{I} \otimes \mathbf{G}^T] \\
&= (\boldsymbol{\Sigma}_K \otimes \cdots \otimes \boldsymbol{\Sigma}_2) \otimes (\mathbf{G}\boldsymbol{\Sigma}_1\mathbf{G}^T).
\end{aligned}$$

Calculation for the covariance of $\mathbf{X} = \mathbf{Y} \times_k \mathbf{G}$ for other values of k proceeds analogously. \square

PROOF OF PROPOSITION 4.1 The density can be obtained as a re-expression of the density of $\mathbf{e} = \text{vec}(\mathbf{E}) = \text{vec}(\mathbf{Y} - \mathbf{M})$, which has a multivariate normal distribution with mean zero and

covariance $\Sigma_K \otimes \cdots \otimes \Sigma_1$. The re-expression is obtained using the following identities,

$$\begin{aligned}
\|(\mathbf{Y} - \mathbf{M}) \times \Sigma^{-1/2}\|^2 &= \langle \mathbf{E}, \mathbf{E} \times \Sigma^{-1} \rangle \\
&= \text{vec}(\mathbf{E})^T \text{vec}(\mathbf{E} \times \Sigma^{-1}) \\
&= \mathbf{e}^T (\Sigma_K \otimes \cdots \otimes \Sigma_1)^{-1} \mathbf{e}, \text{ and} \\
|\Sigma_K \otimes \cdots \otimes \Sigma_1| &= \prod_{k=1}^K |\Sigma_k|^{n_k},
\end{aligned}$$

where $n_k = \prod_{j:j \neq k} m_j$ is the number of columns of $\mathbf{Y}_{(k)}$, i.e. the “sample size” for Σ_k . \square

PROOF OF PROPOSITION 4.3. We first obtain the full conditional distributions for the matrix normal case. Let $\mathbf{Y} \sim \text{anorm}(\mathbf{0}, \Sigma \otimes \Omega)$ and $\Sigma^{-1} = \Psi$. Let (\mathbf{a}, \mathbf{b}) form a partition of the row indices of \mathbf{Y} , and assume the rows of \mathbf{Y} are ordered according to this partition. The quadratic term in the exponent of the density can then be written as

$$\begin{aligned}
\text{tr}(\Omega^{-1} \mathbf{Y}^T \Psi \mathbf{Y}) &= \text{tr} \left(\Omega^{-1} (\mathbf{Y}_a^T \mathbf{Y}_b^T) \begin{pmatrix} \Psi_{aa} & \Psi_{ab} \\ \Psi_{ba} & \Psi_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \right) \\
&= \text{tr}(\Omega^{-1} \mathbf{Y}_a^T \Psi_{aa} \mathbf{Y}_a) + 2\text{tr}(\Omega^{-1} \mathbf{Y}_b^T \Psi_{ba} \mathbf{Y}_a) + \text{tr}(\Omega^{-1} \mathbf{Y}_b^T \Psi_{bb} \mathbf{Y}_b).
\end{aligned}$$

As a function of \mathbf{Y}_b , this is equal to a constant plus the quadratic term of the matrix normal density with row and column covariance matrices of Ψ_{bb}^{-1} and Ω , and a mean of $-\Omega_{bb}^{-1} \Omega_{ba} \mathbf{Y}_a$. Standard results on inverses of partitioned matrices give the row variance as $\Psi_{bb}^{-1} = \Sigma_{bb} - \Sigma_{ba}(\Sigma_{aa})^{-1} \Sigma_{ab} = \Sigma_{b|a}$ and the mean as $-\Omega_{bb}^{-1} \Omega_{ba} \mathbf{Y}_a = \Sigma_{b|a}(\Sigma_{aa})^{-1} \mathbf{Y}_b$. To obtain the result for the array case, note that if $\mathbf{Y} \sim \text{anorm}(\mathbf{0}, \Sigma_1 \circ \cdots \circ \Sigma_K)$ then the distribution of $\mathbf{Y}_{(1)}$ is matrix normal with row covariance Σ_1 and column covariance $\Sigma_K \otimes \cdots \otimes \Sigma_2$. The conditional distribution can then be obtained by applying the result for the matrix normal case to $\mathbf{Y}_{(1)}$ with $\Sigma = \Sigma_1$ and $\Omega = \Sigma_K \otimes \cdots \otimes \Sigma_2$. \square

References

- M. W. Browne and A. Shapiro. Invariance of covariance structures under groups of transformations. *Metrika*, 38(6):345–355, 1991. ISSN 0026-1335. doi: 10.1007/BF02613631. URL <http://dx.doi.org/10.1007/BF02613631>.
- A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981. ISSN 0006-3444. doi: 10.1093/biomet/68.1.265. URL <http://dx.doi.org/10.1093/biomet/68.1.265>.
- L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

- P Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.
- A.T. Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, 23(11):3105–3119, 1994.
- Robert E. Kass and Larry Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, 90(431):928–934, 1995. ISSN 0162-1459.
- Tamara G. Kolda. Multilinear operators for higher-order decompositions. Technical Report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, April 2006.
- Nelson Lu and Dale L. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.*, 73(4):449–457, 2005. ISSN 0167-7152. doi: 10.1016/j.spl.2005.04.020. URL <http://dx.doi.org/10.1016/j.spl.2005.04.020>.
- Matthew W. Mitchell, Marc G. Genton, and Marcia L. Gumpertz. A likelihood ratio test for separability of covariances. *J. Multivariate Anal.*, 97(5):1025–1043, 2006. ISSN 0047-259X. doi: 10.1016/j.jmva.2005.07.005. URL <http://dx.doi.org/10.1016/j.jmva.2005.07.005>.
- Anuradha Roy and Ravindra Khattree. On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. *Stat. Methodol.*, 2(4):297–306, 2005. ISSN 1572-3127. doi: 10.1016/j.stamet.2005.07.003. URL <http://dx.doi.org/10.1016/j.stamet.2005.07.003>.
- Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172, 1984. ISSN 0090-5364. doi: 10.1214/aos/1176346785. URL <http://dx.doi.org/10.1214/aos/1176346785>.
- M.S. Srivastava, T. von Rosen, and D. Von Rosen. Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370, 2008.
- P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001. ISSN 0022-3239. doi: 10.1023/A:1017501703105. URL <http://dx.doi.org/10.1023/A:1017501703105>.
- L. R. Tucker. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, pages 109–127, 1964.
- L.R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.